Do not get lost in projection: finding the right distance for meaningful UMAP embeddings

Eva Blanco-Mallo[1], Verónica Bolón-Canedo[1], and Beatriz Remeseiro[2] *

1- CITIC, Universidade da Coruña, A Coruña, Spain

2- Universidad de Oviedo, Gijón, Spain

Abstract. Dimensionality reduction techniques are essential for visualizing and analyzing high-dimensional data. This study explores the impact of distance measures on the performance of Uniform Manifold Approximation and Projection (UMAP), a widely used dimensionality reduction method. We evaluate their influence on cluster separation, structure preservation, and their effectiveness when used as a preprocessing step for classification tasks on real and synthetic datasets. The results highlight the importance of tailoring distance measures to specific data contexts and provide guidance for optimizing UMAP applications.

1 Introduction

Analyzing and visualizing high-dimensional data is essential in artificial intelligence and data science. Dimensionality reduction techniques simplify complex datasets into lower-dimensional representations, enhancing interpretability, uncovering intrinsic patterns, and improving computational efficiency while mitigating overfitting. Uniform Manifold Approximation and Projection (UMAP) has gained widespread recognition for its ability to preserve both local and global data structures, often surpassing traditional approaches such as principal component analysis and t-distributed stochastic neighbor embedding [1]. Its flexibility and efficiency have enabled diverse applications, including genomics [2], natural language processing [3], cheminformatics [4], and healthcare [5], where UMAP facilitates intuitive exploration of complex patterns and supports critical decision-making processes.

Despite its advantages, UMAP performance is highly sensitive to hyperparameter selection. Among these, the distance measure plays a critical role, directly influencing clustering, outlier detection, and the interpretability of embeddings. However, the default Euclidean distance is often used without investigating potentially more effective alternatives. Few studies have specifically

*Grants PID2019-109238GB-C22, PID2023-147404OB-I00, TED2021-130599A-I00, PID2021-128045OA-I00, and FPI PRE2020-092608 funded by MICIU/AEI/10.13039/501100011033. Grant GRU-GIC-24-018 funded by the Agency for Science, Business Competitiveness, and Innovation of the Principality of Asturias in Spain (SEKUENS). CITIC, as a center accredited for excellence within the Galician University System and a member of the CIGUS Network, receives subsidies from the Department of Education, Science, Universities, and Vocational Training of the Xunta de Galicia. Additionally, it is co-financed by the EU through the FEDER Galicia 2021-27 operational program (Ref. ED431G 2023/01). Grant ED431C 2022/44 funded by Xunta de Galicia.

addressed the influence of distance measures on UMAP performance. For example, Smets et al. [6] and Vermeulen et al. [7] demonstrated that cosine distance improves UMAP performance in mass spectrometry imaging when combined with optimized hyperparameters. However, these findings are heavily context-specific, and a systematic understanding of how distance measures affect UMAP behavior across broader applications remains lacking.

This study fills this gap by assessing the performance of five commonly used distance measures. Using synthetic and real-world datasets, we examine their impact on cluster separation, structure preservation, and the separability of classes in the resulting embeddings for downstream classification tasks. The findings offer practical recommendations for selecting the most suitable distance measure based on dataset characteristics and application requirements.

2 Materials and methods

Uniform Manifold Approximation and Projection. UMAP is a manifold learning technique designed to preserve the topological structure of high-dimensional data by constructing a graph of relationships between data points and optimizing it in a lower-dimensional space. Its performance depends on three critical hyperparameters: the number of neighbors (n), the minimum distance (min_dist), and the distance measure. The hyparameter n balances local and global structure preservation, with smaller values focusing on local patterns and larger values capturing broader relationships. The min_dist controls the density of points in the reduced space, where lower values create tighter clusters and higher values produce more dispersed embeddings. The distance measure determines how similarities between data points are computed in the high-dimensional space, directly impacting clustering, outlier detection, and the overall interpretability of the embeddings. While Euclidean distance is the default, other metrics may better align with specific data characteristics.

Distance measures. In this study, five widely used distance measures are evaluated: Euclidean (ED), Manhattan (MD), cosine (COD), Canberra (CAD), and Bray-Curtis (BD). ED computes the straight-line distance between two points, assuming isotropy and consistent geometric relationships. MD computes the sum of absolute differences across dimensions, capturing variability independently in each feature. COD evaluates the cosine of the angle between two vectors, emphasizing directional similarity rather than magnitude. CAD computes normalized differences along each dimension by dividing the absolute difference by the sum of paired values, making it particularly sensitive to small magnitudes and emphasizing subtle variations in feature values. BD measures compositional dissimilarity by comparing proportional differences across dimensions, rather than absolute magnitudes. The intrinsic properties of each metric determine UMAP's ability to adapt to various data structures, offering unique advantages depending on the dataset.

Datasets. To evaluate UMAP's ability to preserve geometric and structural relationships, we utilized three widely recognized synthetic datasets: Two

Lines, Two Circles, and Two Moons (Figure 1, left column). These benchmarks, extensively used for testing clustering and dimensionality reduction techniques, present unique structural challenges. The Two Lines dataset, composed of parallel lines, tests the effectiveness of UMAP in maintaining linear separations. Two Circles, with its concentric arrangement, evaluates the algorithm's capacity to distinguish nested structures. Two Moons, with interspersed semicircles, highlights UMAP's ability to handle nonlinear separations. These datasets enable controlled experimentation and provide insights into the influence of distance measures on embedding quality. For real-world scenarios, we selected datasets that reflect diverse structures and practical applications. Word2Vec embeddings, which encode semantic relationships based on word co-occurrence, tested UMAP's ability to retain both global and local semantic coherence, essential for applications in natural language processing and text analytics. Regarding classification tasks, four benchmark datasets were chosen: Wine, Breast Cancer, Digits, and Low Resolution Spectrometer. These datasets encompass varied feature types, dimensionalities, and class distributions. By combining these datasets, our experiments capture UMAP's adaptability across distinct domains, providing actionable insights into the role of distance measures in embedding quality and downstream classification performance.

3 Experimentation

This study provides a comprehensive evaluation of UMAP performance in various data scenarios, from synthetic benchmarks to real-world datasets. The implementation was based on the official UMAP library [8], with datasets sourced from the Scikit-Learn library [9] and the UCI Machine Learning Repository [10]. This setup ensures reproducibility and broad applicability of the results. Due to space limitations, only the results obtained with the default parameter settings ($n = 15$, $min\_dist = 0.1$) are shown in the figures. The complete results and the source code are available in our GitHub repository[1].

Visualization performance on synthetic data. Across the synthetic datasets (Figure 1), MD and BD consistently outperformed other metrics in preserving both linear and nonlinear structures while offering computational efficiency. In the Two Lines and Two Circles datasets, these distances maintained clear cluster separability and preserving the circular structure of the data, even with lower n values, making them ideal for large datasets where reducing computational cost is crucial. In the Two Moons dataset, ED and MD distances excelled, preserving class separability and capturing nonlinear structures effectively, particularly at lower n values. BD also performed well but exhibited distortions at higher n values, thus being more sensitive to the influence of hyperparameters. COD and CAD failed across all datasets, frequently merging classes or distorting structural relationships. These findings highlight MD as the most reliable distance measure for preserving structure and optimizing computational efficiency, followed by ED and BD, although requiring a more refined

[1] `https://github.com/evablanco/distances_analysis_UMAP`

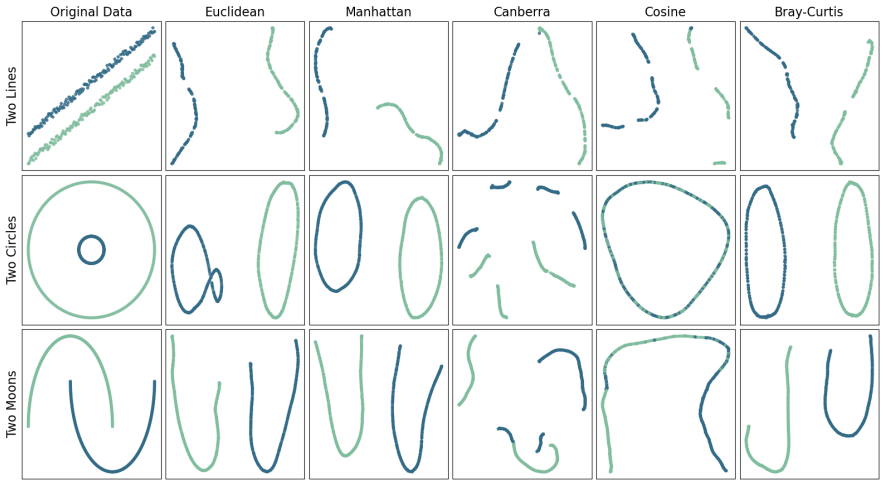

Fig. 1: UMAP embeddings for synthetic datasets with default parameters.

hyperparameter setting.

Visualization performance on real data. The top 10 most similar terms to the following words were obtained using the embeddings generated with Word2Vec: "dream", "paris", "car", and "computer". BD demonstrates the most balanced performance, preserving local cohesion while maintaining effective inter-class separation by leveraging proportional differences. CAD shares these strengths but is less robust to hyperparameter variations, introducing inconsistencies. COD prioritizes directional similarity, yielding tightly packed intra-class clusters with strong global separability, although it can obscure finer distinctions within clusters. In contrast, ED and MD produce more diffuse clusters with weaker inter-class separation, reflecting their reliance on absolute magnitudes over proportional or directional relationships. These results establish BD as the most suitable metric for embedding tasks requiring a balance of local detail and global separation, with CD offering superior performance when compact intra-class clustering and global separability are paramount. Figure 2 illustrates the performance comparison between ED, the default metric, and BD.

Classification performance The support vector machine algorithm was chosen due to its known effectiveness in high-dimensional spaces and its ability to handle nonlinear relationships through the use of different kernels, thereby capturing complex relationships in the data. The accuracy of the model was evaluated both with raw data and with UMAP embeddings, generating representations with 20%, 30%, and 50% of the original feature sizes of the datasets. Table 1 reports the best accuracy values obtained with each distance metric and hyperparameter configuration after 10 repetitions, where $\#c$ indicates the number of components used.

The Wine dataset, characterized by its continuous features such as alcohol content or magnesium levels, requires distance measures that effectively cap-

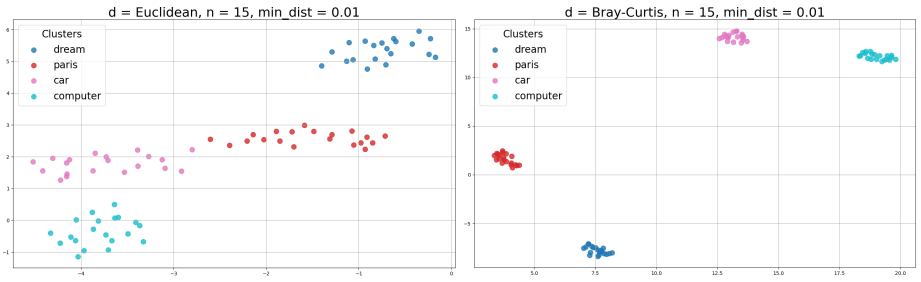

Fig. 2: Comparison of UMAP embeddings from Word2Vec using ED (default metric) and BD (recommended for balancing intra- and inter-class separation).

Table 1: Classification accuracy and best UMAP hyperparameter configurations.

Dataset	Raw		Euclidean		Manhattan		Cosine		Canberra		Bray-Curtis	
	Acc.	Feats.	Acc.	n/min_d/#c	Acc.	n/min_d/#c	Acc.	n/min_d/#c	Acc.	n/min_d/#c	Acc.	n/min_d/#c
Wine	0.58	13	0.64	30/1.0/6	**0.72**	100/0.1/2	0.56	50/0.1/2	0.44	30/0.1/2	0.67	30/0.1/2
Breast C.	0.91	30	**0.95**	100/1.0/6	**0.95**	15/1.0/6	0.89	30/0.01/9	0.86	50/0.1/6	0.91	15/0.01/9
Digits	**0.99**	64	0.30	15/0.01/12	0.32	30/0.1/19	0.28	15/0.01/12	0.22	15/0.1/32	0.31	30/0.1/32
LRS	**0.89**	100	0.53	15/0.01/20	0.55	15/0.1/30	0.62	15/1.0/30	0.55	50/0.1/20	0.63	30/0.01/30

ture both the absolute magnitudes and relationships among features for accurate classification. MD achieved the highest accuracy improvement (+13.89%), reducing the original 13 components to 2 by effectively preserving absolute differences across dimensions. BD follows (+8.34%), benefiting from its focus on proportional differences that effectively capture feature interdependencies while preserving meaningful separations in the embedding space. ED offers better performance than using raw data (+5.56%), but to a lesser extent and requiring more components than the previous ones, probably due to its sensitivity to scale and outliers. In contrast, COD and CAD perform worse than the raw data. COD emphasizes directional similarity, disregarding the absolute magnitudes crucial for class separability in Wine, while CAD's sensitivity to small values introduces distortions by overemphasizing less relevant feature variations.

In the Breast Cancer dataset, where absolute differences between features are also important but values are normalized, both ED and MD achieved the highest accuracy (+3.77%) using only 6 components. BD maintained the original accuracy while reducing components to 9, indicating effective dimensionality reduction without loss of performance. COD and CAD yielded slightly lower accuracies, but significantly reduced the number of components.

Finally, in the Digits and Low Resolution Spectrometer (LRS), a decrease in performance is observed when applying UMAP. This reveal a key limitation of UMAP: its inability to preserve geospatial or sequential dependencies critical in high-dimensional data. While Bray-Curtis (BD) performs better than other metrics, it still falls short, highlighting the need for distance measures that explicitly account for spatial or sequential structures.

4 Conclusion

This study highlights the importance of selecting appropriate distance metrics in UMAP applications based on dataset characteristics and task requirements. For data visualization, the results suggest that MD is the most suitable metric for UMAP applications that require a faithful representation of data structures. ED and BD also provide good results, but require further hyperparameter tuning. These distances effectively preserve class separability and maintain structural integrity in synthetic datasets with linear or simple nonlinear relationships. In natural language processing tasks, where preserving semantic relationships in embeddings is essential, BD and COD are particularly effective. BD excels in capturing intra-class dissimilarities, while COD generates tighter embedding clusters, emphasizing global rather than local relationships. For dimensionality reduction in classification tasks, MD emerges as the most effective measure, achieving significant accuracy improvements while substantially reducing dimensionality in datasets where absolute differences between magnitudes are critical. However, in high-dimensional datasets characterized by geospatial or sequential dependencies, UMAP is not recommended as a preprocessing step for classification with any distance measure, as it fails to preserve these critical spatial dependencies. This work represents a first step toward understanding the influence of distance metrics on UMAP performance. A more comprehensive analysis incorporating a broader range of datasets, including those with higher complexity and varied characteristics, is required to generalize these findings and further refine UMAP's applicability across diverse data contexts.

References

[1] Ghojogh et al. *Uniform Manifold Approximation and Projection (UMAP)*, pages 479–497. 2023.

[2] Dorrity et al. Dimensionality reduction by umap to visualize physical and genetic interactions. *Nature Communications*, 11(1):1537, 2020.

[3] David Silva and Fernando Bação. Mapintel: Enhancing competitive intelligence acquisition through embeddings and visual analytics. In *EPIA Conference on Artificial Intelligence*, pages 599–610. Springer, 2022.

[4] Humer et al. Cheminformatics model explorer (cime): exploratory analysis of chemical model explanations. *Journal of Cheminformatics*, 14(1):21, 2022.

[5] Abdullah et al. Visual analytics for dimension reduction and cluster analysis of high dimensional electronic health records. In *Informatics*, volume 7, page 17. MDPI, 2020.

[6] Smets et al. Evaluation of distance metrics and spatial autocorrelation in uniform manifold approximation and projection applied to mass spectrometry imaging data. *Analytical chemistry*, 91(9):5706–5714, 2019.

[7] Vermeulen et al. Application of uniform manifold approximation and projection (umap) in spectral imaging of artworks. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 252:119547, 2021.

[8] McInnes et al. Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29):861, 2018.

[9] Pedregosa et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[10] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.